

Exploring the Mechanism of Tonal Contraction in Taiwan Mandarin

Chierh Cheng¹, Yi Xu¹, Michele Gubian²

¹Department of Speech, Hearing and Phonetic Sciences, University College London, UK

²Centre for Language & Speech Technology, Radboud University, Nijmegen, NL

Chierh.Cheng@googlemail.com, Yi.Xu@ucl.ac.uk, M.Gubian@let.ru.nl

Abstract

This study investigates the mechanism of tonal contraction when a disyllabic unit is merged into a monosyllable at fast speech rate in Taiwan Mandarin. Various degrees of contraction of bi-tonal sequences were elicited by manipulating speech rates. Functional Data Analysis was performed to compare trajectories of F_0 and velocity in the contracted and non-contracted syllables. Preliminary results show that speakers always make an effort to produce the original tones, even in cases of extreme degrees of reduction. This finding militates against phonology-based accounts like the *Edge-in* model, according to which contraction is a process of deleting adjacent tonemes while leaving the non-adjacent tonemes intact.

Index Terms: tone, contraction, Taiwan Mandarin, Functional Data Analysis, Edge-in model

1. Introduction

Phonetic reduction of a sequence of two or more syllables into one has been noted to be pervasive in the Sinitic language family, and it has been referred to as ‘syllable contraction’, ‘syllable merger’ or ‘syllable fusion’ [6, 17, 18]. The goal of this work is to analyze tones in syllable contraction. Taiwan Mandarin, as standard Chinese spoken in Taiwan, has four lexical tones: High (H, 55), Low (L, 21 or 213 if occurs prepausally), Rising (R, 35) and Falling (F, 51). Digits in parenthesis are conventional numeric notation for tonemes, also as schematized in Figure 1.

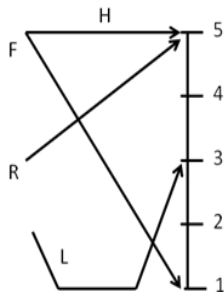


Figure 1: The conventional five-scale tone representation scheme proposed by Chao [2]. The digit 5 represents the highest pitch value while 1 the lowest.

Formal phonology with the contribution of morphology of Chinese has proposed various derivational rules to predict the output values of a contracted syllable from their underlying citation elements, which operates in an *Edge-in* fashion as first suggested by Yip [21]. That is, the output tone is composed of the two edge tonemes of the source syllables. Figure 2 shows an *Edge-in* process.

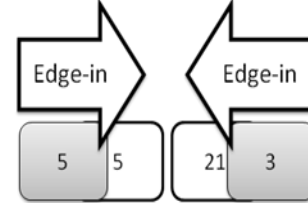


Figure 2: An *Edge-in* approach to derive the output tone 53 from two source syllables, ex. /kən55/ + /pən213/ → /kən53/, means “basically” (the bilabial plosive /p/ gives rise to a realisation of coda /m/).

Because no general consensus has been reached on which tonal elements are deleted or preserved in the contracted output, other accounts such as the Sonority model [11] and Optimality Theory [10] were also adapted to improve the original *Edge-in* model. Recent experimental studies, however, show that exceptions to this formal generalization are not uncommon. In general, phonetic-oriented papers consistently report that, whichever the underlying tonal combinations are, flat or only slightly sloping contours are the typical F_0 contours in the contracted syllables [3, 5, 12].

None of the previous accounts, however, has considered the possibility that contraction is only a severe form of phonetic undershoot and the underlying tonal elements are reduced rather than completely deleted. We hence propose the following hypothesis: when contraction occurs, speakers still try to approach each underlying tonal target. In particular, in case of a disyllable contracted into a monosyllable, traces of the two source tones are still detectable in F_0 trajectories despite being squeezed into a single syllable. To test this hypothesis, we need to solve two problems: how to elicit reduction in a systematic way and how to qualitatively and quantitatively determine whether the underlying tonal elements are present.

In regard to the first problem, several studies have indicated that speech rate is a reliable predictor of the amount and degree of reduction [4, 22]. With a set of balanced and well-designed stimuli, we expect to elicit contracted syllables simply by asking subjects to speed up their speech production in the laboratory. In regard to the second problem, i.e. how to produce evidence on whether or not tonal elements are still present in the contracted forms, we perform two kinds of comparison. One is to compare the same tone combination (e.g. HR) in the same preceding and following tone context in *contracted* and *non-contracted* modes. The other is to compare a *contracted* combination (e.g. HR) with a single *non-contracted* tone, again in the same preceding and following context. The single *non-contracted* tone is chosen according to the *Edge-in* model, which predicts the surface form of tone combinations in case of contraction. For example, when a HR sequence is reduced to only one syllable (5535 → 55), the *Edge-in* model predicts a surface form of H (55). In order to perform this analysis quantitatively, a functional t-test, a tool within the

Functional Data Analysis framework [15], will be provided for each comparison. In this way, the time intervals where the difference between two average curves is significant can be seen. Additionally, average F_0 velocity will be compared as well, since they capture well the articulatory movements toward the underlying tone targets [9].

We will show that (i) *contracted* and *non-contracted* modes tend to contain the same gestures such as rises and falls in the same order. On the other hand, (ii) *contracted* sequences appear to differ in a qualitative way from their respective *non-contracted* tone (according to *Edge-in*). Both these findings will help to support our hypothesis that tonal targets are retained even in severe reduction.

2. Experiment

2.1. Stimuli

Disyllabic /ma+/ma/ nonsense sequence with a total of 16 (4x4) tonal combinations embedded in two carrier sentences were constructed for testing materials. The segments of target sequence were chosen to be /ma+/ma/, so as to facilitate segmentation and better F_0 inspection (媽麻馬罵 for tones H R L F as in Chinese characters, respectively). Table 1 lists the two carrier sentences, differing in the tones preceding the target sequences, i.e. H and L. Note that the tone following the target sequences is always H. Each carrier sentence consists of three phrases. The first phrase consists of 9 underlying syllables, the second 13, and the third 17. This helps to impose different amounts of time pressure on the target disyllabic sequences. The same target sequence was embedded in each phrase and hence produced three times within each carrier sentence.

Table 1. *Stimuli and carrier sentences.*

Carrier sentences with a high/low preceding tone	
Characters	你想吃/買 ____ 沙拉是吧！我當然不吃/買 ____ 沙拉那種東西，因為我不喜歡/欣賞 ____ 沙拉那種酸酸的醬料！
Pinyin	ni xiang chiH/maiL ____ shaH la shi ba! wo dang ren bu chiH/maiL ____ shaH la na zhong dong xi, yin wei wo bu xi huanH/xin shangL ____ shaH la na zhong suan suan de jiang liao!
English	You want to eat/buy ____ salad, didn't you! Of course I won't eat/buy ____ salad that kinda stuff, because I dislike the sour source of ____ salad.

2.2. Recordings

Six male Taiwan Mandarin speakers were recorded. The speakers were aged between 21 and 28 and had neither self-reported speech disorders nor professional vocal training. The recordings were conducted in the anechoic chamber of University College London. Speech was recorded with a Shure SM10A microphone placed approximately 30 centimeters from the subjects' mouth. All stimuli were presented to the subjects in traditional Chinese characters and each time only one carrier sentence with the embedded stimuli was shown on the screen in front of the seated subject. To control the level of time pressure, subjects were instructed to articulate the material at three speaking rates, *slow/clear* as if reciting in class, *natural* as if having a conversation with a friend, and as *fast* as possible. The exact speed of articulation, however, was left to the subjects' own discretion. The average speech rates of slow, natural and fast speech across the six subjects were 4.5, 6.1 and 9.3 underlying syllables per second, respectively. In this

fashion, we aim to obtain both canonical and elliptic forms of each target sequence for every position within the carrier sentence and rate of speech. Three randomized blocks of the above 16 sentence sequences were used. In total, the number of target sequences for analysis was 16 (stimuli) \times 6 (subjects) \times 3 (positions in the carrier) \times 2 (preceding tones) \times 3 (speech rates) \times 3 (blocks/repetitions) = 5184 tokens. With this systematically-designed experiment, the large dataset has been collected not only restricted to the current study but also for a long-term investigation.

2.3. Measurements

All sound files were segmented by the first author, a native Taiwan Mandarin speaker. First the target /ma+/ma/ tokens were isolated from the rest of the material. Then a boundary between the two syllables was marked whenever a clear second nasal was found, the token being labeled as *non-contracted*. When the second nasal is absent no boundary was marked and the token was labeled as *contracted*. Intermediate cases were excluded from this analysis.

The extraction of F_0 contours was first done with the vocal cycle marking of the Praat programme and then with manual repair of octave jumps and other apparent irregularities [1]. A Praat script was used to convert the vocal period into F_0 values, and then to smooth the resulting curves using a trimming algorithm that eliminated sharp bumps and edges [20]. For each target curve, 40 points of measurement were generated, 20 equidistant points per syllable for *non-contracted* tokens and 40 equidistant points for *contracted* tokens.

2.4. Functional Data Analysis

Functional Data Analysis (FDA) refers to a set of tools that extend ordinary multivariate statistics to the domain of functions. In this work, the FDA framework is used to compare F_0 contours that have different durations in a principled way. A standard data preparation procedure has to be followed (for details see [15] or the website maintained by the third author¹). All sampled F_0 contours were converted to semitones and their average through time was subtracted. This helps the automatic extraction of shape-related features. Then all sampled curves have to be interpolated using the same function basis (B-splines in our case). The use of a common function basis forces all curves to be defined on a common time interval. Since the original F_0 contours have different durations, their functional representations have to be distorted (registered) in time in such a way that they exhibit the same duration. To make this process less detrimental, points that have the same meaning across the curve set are used as landmarks and get automatically aligned in time across the registered functions set. We aligned the boundaries between the first and the second syllable wherever present.

Comparisons between pairs of F_0 contours sets are carried out as follows. The average function is extracted from each set of curves, i.e. a function whose value at every point in time is the average of all the functions in the set at that same time (e.g. Figure 3a). The same goes for F_0 velocity profiles, which are calculated before the time registration, thus the original velocity values are preserved. Functional t-test is also applied to the groups of functions whose averages are displayed (e.g. Figure 3b). Functional t-test extends the idea of t-test by showing the significance of the difference between the means of two groups of functions at each instant in time. All FDA operations were carried out using the freely available R package 'fda' [14, 16].

3. Results and Discussion

3.1. Comparison between Non-contracted and Contracted HR curves

Analysis of the above experiment for the HR tone combination is presented here to illustrate the difference between the *contracted* and *non-contracted* modes. To keep the carry-over tonal influence from the preceding tonal context constant, HR curves with different preceding tones were analysed separately: H#HR (HR following an H shown in Figure 3) and L#HR (HR following an L), the latter not being shown here. Figure 3 consists of 4 plots: (a) The normalised average F_0 contours of *contracted* and *non-contracted* HR as a function of time (note that the average F_0 has been removed as mentioned in 2.4), (b) the corresponding functional t-test on F_0 contours between these two modes, (c) the normalised average velocity of *contracted* and *non-contracted* HR as a function of time and (d) the corresponding functional t-test on velocity between these two modes. In (a) and (c) the green thick lines are averages of 125 *non-contracted* HR curves, while the black lines are averages of 24 *contracted* HR curves. In the t-tests plots, i.e. (b) and (d), the dashed horizontal blue line gives the 0.05 critical value for the t-statistic and the red solid line represents the observed statistic. In practice, in intervals where the solid red line lies above the dashed horizontal line, the difference between the corresponding average functions is significant. Note that the sudden decrease towards minima seen in the red line in many of the t-test plots is simply a consequence of the respective average curves crossing each other. When this occurs the differences are of course small, and consequently (likely to be) insignificant.

For an underlying HR we expect to see a high level tone followed by a downward movement which precedes a final rise. It is noted that when H is followed by R, F_0 needs to first move down (negative velocity) in order to execute a rising pattern (positive velocity). In Figure 3a, *non-contracted* HR shows such a canonical trajectory while *contracted* HR keeps sloping downwards till the end of the unit without this final tonal rise. In their velocity profiles (Figure 3c), both modes first decrease towards extrema and attain negative values (corresponding to a decreasing F_0 contour) and then they *both* change direction beginning to rise by increasing velocity. This trend is interpreted as revealing the underlying intentions of the speakers to approach the R target by crossing the zero-velocity line. That is, our first comparison shows that speakers still attempt to approach each underlying tonal target with an extensive movement in the *non-contracted* units, thus the same basic movements arise in both modes. The absence of this final rise in *contracted* HR in Figure 3a can be accounted for by the shorter duration in contracted syllables which prevents the velocity change to be translated into substantive changes in the overall F_0 contours.

Figure 3b shows that *non-contracted* and *contracted* HR F_0 contours differ significantly almost everywhere (except at the inevitable crossing points). However, Figure 3a shows that those differences are mainly due to different amplitude of the same gestures. Figure 3d indicates that velocity profiles differ mainly in the second half, where the R target is achieved in *non-contracted* tokens but only attempted in *contracted* ones. More specifically, the two red peaks in Figure 3d reflect two regions of difference in the velocity of syllable 2, steeper negative velocity in the first half of the *non-contracted* syllable 2 and steeper positive velocity in the second half of the *non-contracted* syllable 2. These two peaks are consistent with the Fujisaki model in which a dynamic tone is represented by two adjacent commands in opposite directions [7]. In the Target Approximation model, the same results can be predicted with

one dynamic command [13].

3.2. Comparison between Contracted HR and Single H

In this section we compare *contracted* HR combinations against single *non-contracted* H tones. According to the *Edge-in* model, HR should be realised as a high level tone H when contracted into a single syllable, since its source edge tonemes are both high.

Figure 4 shows the comparison between *contracted* HR curves and a single H tone in the same tonal environment, i.e. both are surrounded by H tones. In (a) and (c) the green thick lines are averages of 119 *non-contracted* H curves, while the black lines are averages of 24 *contracted* HR curves (the same as before). In Figure 4a-b, the significant difference in the left half likely reflects the effect of anticipatory dissimilation [8, 19], with which the F_0 of the preceding H is raised by the following R. In Figure 4c-d, the significant difference in the right half reflects the presence/absence of underlying R. Here the comparison again demonstrates that speakers still attempt to achieve the underlying targets within a limited duration and do not delete certain items or preserve others as proposed by the *Edge-in* model.

Thus, the empirical phonetic data from the present study does not appear to conform to the predictions made by the *Edge-in* model regarding tones in contracted syllables. Other tone combinations in different tonal contexts have also been analysed and the results of these combinations are largely in line with what is shown in this and in the previous section. (An extended analysis is available online¹).

4. Conclusions

In this paper, two kinds of comparison were performed to test the hypothesis whether speakers try to approach each underlying tonal target even when contraction occurs. Results from both comparisons seem to support this hypothesis. With the support of F_0 velocity contours and of functional t-test, we have found evidence that even when a two-syllable sequence is reduced to one syllable, the underlying source tonemes are still present. This current study furthers our understandings by showing that the tonal values in a contracted output are not necessarily derived by rules as proposed by the *Edge-in* model, and the nature of the surface variation seems to be articulatorily rather than phonologically driven.

5. Acknowledgements

The research of Michele Gubian is supported by the Marie Curie Research Training Network Sound-to-Sense (<http://www.sound2sense.eu>). We would like to thank Lou Boves for his time and advice on this paper.

6. Endnotes

¹ Online: <http://lands.let.ru.nl/FDA/>

7. References

- [1] Boersma, P. & Weenink, D., “Praat: doing phonetics by computer [Computer program]”, Version 5.1.32, retrieved 30 April 2010 from <http://www.praat.org/>, 2010.
- [2] Chao, Y. R., “Mandarin primer, an intensive course in spoken Chinese”, Cambridge: Harvard University Press, 1948.
- [3] Cheng, C. E., “An acoustic phonetic analysis of tone contraction in Taiwan Mandarin”, MA, National Cheng Chi University, Taipei, 2004.

- [4] Cheng, C. & Xu, Y., “Extreme reductions: Contraction of disyllables into monosyllables in Taiwan Mandarin”, *INTERSPEECH*, 456-459, 2009.
- [5] Chung, K. S., “Contraction and backgrounding in Taiwan Mandarin”, *Concetric: Studies in Linguistics*, 32: 69-88, 2006.
- [6] Chung, R. F., “Syllable contraction in Chinese”, *Chinese Languages and Linguistics III: Morphology and Lexicon*, Academia Sinica, Taipei, 199-235, 1997.
- [7] Fujisaki, H., Wang, C., Ohno, S. and Gu, W., “Analysis and synthesis of fundamental frequency contours of standard Chinese using the command-response model”, *SPEECH COMMUN.*, 47:59-70, 2005.
- [8] Gandour, J., Potisuk, S. & Dechongkit, S., “Tonal coarticulation in Thai”, *J. PHONETICS*, 22: 477-492, 1994.
- [9] Gauthier, B., Shi, R. S., & Xu, Y., “Learning phonetic categories by tracking movements”, *Cognition*, 103: 80-106, 2007.
- [10] Hsiao, Y.-C. E., “Tone contraction”, *Chinese languages and linguistics VIII*, 1-16, Taipei: Academia Sinica, 2002.
- [11] Hsu, H.C., “A Sonority Model of Syllable Contraction in Taiwanese Southern Min”, *Journal of East Asian Linguistics*, 12: 349-377, 2003.
- [12] Myers, J. & Li, Y. S., “Lexical frequency effects in Taiwan Southern Min syllable contraction”, *J. PHONETICS*, 37: 212-230, 2009.
- [13] Prom-on, S., Xu, Y. and Thipakorn, B., “Modeling tone and intonation in Mandarin and English as a process of target approximation”, *J. ACOUST. SOC. AM.*, 125(1):405-424, 2009.
- [14] R Development Core Team, “R: A language and environment for statistical computing,” R Foundation for Statistical Computing, Vienna, Austria, URL: <http://www.R-project.org>, 2008.
- [15] Ramsay, J. O. & Silverman, B. W., “Functional Data Analysis—2nd Ed”, Springer, 2005.
- [16] Ramsay, J. O., Hookers, G. & Graves, S., “Functional Data Analysis with R and MATLAB”, Springer, 2009.
- [17] Tseng, S. C., “Monosyllabic word merger in Mandarin”, *LANG. VAR. CHANGE*, 17: 231-256, 2005.
- [18] Wong, W. Y. P., “Syllable fusion in Hong Kong Cantonese connected speech”, Ph.D., Ohio State University, 2006.
- [19] Xu, Y., “Contextual tonal variations in Mandarin”, *J. PHONETICS*, 25: 61-83, 1997.
- [20] Xu, Y., “_ProsodyPro.praat”, 2005-2010. Online: <http://www.phon.ucl.ac.uk/home/yi/tools.html>, accessed on 30 April 2010.
- [21] Yip, M., “Template morphology and the direction of association”, *Natural Language and Linguistic Theory*, 6: 551-577, 1988.
- [22] Yu, A. C. L., “Understanding near mergers: The case of morphological tone in Cantonese”, *Phonology*, 24:187-214, 2007.

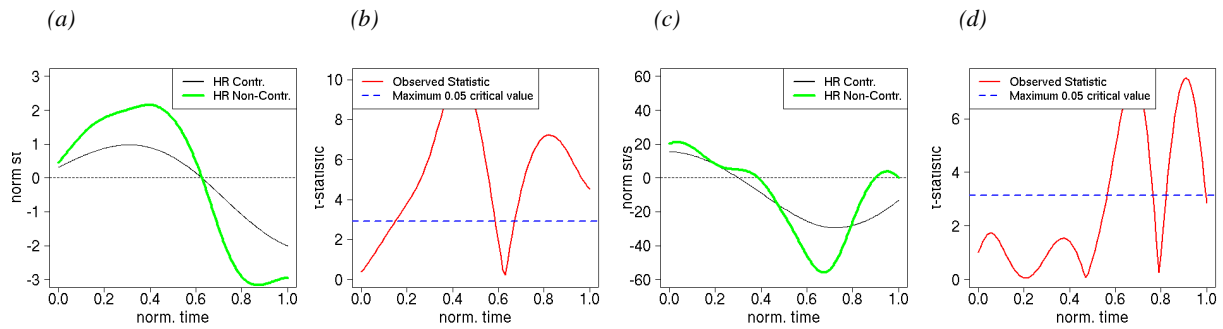


Figure 3: Contracted HR versus non-contracted HR, (H preceding tone, H#HR).

From left to right, (a) Normalised average F_0 contours of contracted HR (black) and non-contracted HR (thick green), (b) Functional t-test on normalised average F_0 contours, (c) Normalised average F_0 velocity contours and (d) Functional t-test on normalised average F_0 velocity contours.

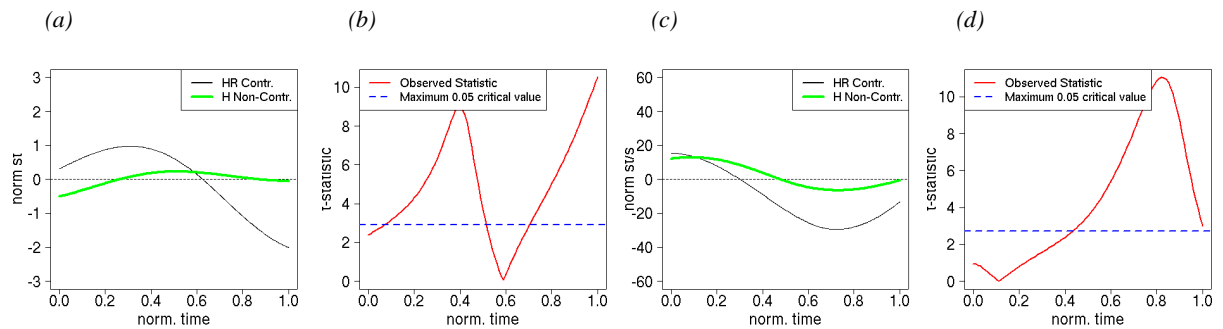


Figure 4: Contracted HR versus single H from non-contracted HH (Contr: H#HR#H versus Non-Contr: H#HH#H).

From left to right, (a) Normalised average F_0 contours of contracted HR (black) and single H (thick green), (b) Functional t-test on normalised average F_0 contours, (c) Normalised average F_0 velocity contours, and (d) Functional t-test on normalised average F_0 velocity contours.